IT UNIVERSITY OF COPENHAGEN

Data in the Wild: Wrangling and Visualising Data (Autumn 2022)

Course code: KSDWWVD1KU

By

Resit Kadir, reka@itu.dk

Bence Kovács, bkov@itu.dk

Hans Christian Traugott-Olsen, hatr@itu.dk

Identifying drivers of property prices in Copenhagen and Greater Copenhagen area

1st Kadir, 2nd Kovács, 3rd Traugott-Olsen
IT University of Copenhagen, Denmark
reka@itu.dk, bkov@itu.dk, hatr@itu.dk

***Abstract*—News report on property prices in Copenhagen reaching new highs every year. Can you even afford anything, many think to themselves? What if you lived near, but outside Copenhagen? We set out to understand the property prices in the Copenhagen and Greater Copenhagen area. What drives the prices? Are the properties cheaper outside of Copenhagen? We scraped Danish property listing websites to better understand these questions, using visualizations and regression analysis. We found that the energy label plays an important role, as well as the location. Generally, the more we move out of Copenhagen, the cheaper it is (excluding northern parts such as Hellerup). Additionally, we created an interactive map that can be used to easily investigate listed properties and their respective information (size, prize, location, etc).**

I. Introduction

Many life forms find a place that is to be their sanctuary. For humans, this is called a home. Historically, humans would either build their own home, or buy one at an affordable price. Nowadays, however, property prices have been on the rise since the 2007 housing bubble, and have even hit historical highs, with apartments being 4.5 times as expensive, and houses 3.7 times as expensive as in 1992, in the Greater Copenhagen and Copenhagen area [1].With house prices increasing, so too does the financial impact of the decision. It is therefore important that young people try to optimize this decision, to not put themselves financially behind from the beginning. But what factors should one be aware of when looking for a property in Copenhagen Greater Copenhagen? To answer this, we set our research question to be:

What are the main drivers of property prices in Copenhagen Greater Copenhagen area?

II. Scraping

Web scraping is a good method to use when there is no alternative structured data set available [1]. It makes the impossible task of manually going through all listed properties and manually recording the data possible, by automating the whole process, without disrupting the site when done ethically. We automatically built a structured, but raw, data set using this method.

While scraping fit our circumstances, it did come with its drawbacks. Using this method did not allow us to build a data set for the whole of Denmark. It was simply too slow when scaling up, especially if it was not perfectly optimized, and our computing power was limited to an ordinary laptop. As a result, even with smaller operations, such as our limited scope of Copenhagen / Greater Copenhagen, it took more than six hours for the scraper to run. That being said, scraping allowed us to relatively quickly accommodate the frequent change in listed properties as one might be sold / put up for sale at any given time.

Before scraping any website, we made sure to read the robot.txt file, and we checked with the Danish regulation on web scraping to make sure we did not scrape unethically [2].

In order to narrow the search down to Copenhagen and Greater Copenhagen, we had to build our own dictionary of zipcodes that relate to a city or city part of the specified area. The dictionary was built manually by inspecting google maps. Google maps has purple borders for Copenhagen and it's respective areas. From those borders, the zipcodes of each city and city part were recorded into the dictionary. The scraper then used this dictionary to guide it's scraping.

A. Boliga.dk and Boligsiden.dk

We were able to able to gather information on properties in Copenhagen and Greater Copenhagen in the following manner: Firstly, we decided to use Boliga.dk to gather information from the Copenhagen area, and Boligsiden.dk to gather information from Greater Copenhagen area - Data was scraped in the period from October 13-November 21.

The scraper was built using the web-based automation tool Selenium. Boliga.dk was navigated by us, to understand which data we needed and how it was obtained through inspecting the html of the website. We decided to scrape all data about the properties, except the written description. In order to get the data from the Greater Copenhagen area, we repurposed the scraper to be compatible with boligsiden.dk. This was easily done, as property sites are oddly enough completely streamlined.

In the end, we ended up gathering 2214 properties from boliga.dk in the Copenhagen area, and 4016 in the Greater Copenhagen area from boligsiden.dk.

[1] https://rkr.statistikbank.dk/201

[2] www.hjulmandkaptain.dk/nyheder/er-dataskrab-lovligt

After both websites were scraped, we merged the final data set:

TABLE I
MERGED PROPERTY DATA

	Data		
	Boliga.dk	***Boligsiden.dk***	***Merged***
Observations	2214	4016	6230

[a]Individual and merged observations from Boliga.dk and Boligsiden.dk.

B. Scraping nearby Restaurants, day care facilities, schools and the like

We hypothesized that having great facilities, restaurants and the like would impact the price of a property, as the demand for such would be higher. Therefore, we used the library GeoPy 2.3.0 [3] to obtain the coordinates of properties by running the address through the library. The idea was to create areas using the coordinates and then count the amount of restaurants, institutions etc. in that area. However, Google Maps did not let us filter for a given area, using this approach. Instead we tried to scrape it by using the addresses obtained from the website scraping and then added "Nearby restaurants" to it, which worked when done manually.

We were unable to make this part of the scraper work. Google Maps would give us restaurants, institutions, etc. that were very far away from the queried address (sometimes many kilometers). We were not able to find out why this was sometimes the case with the scraper, as opposed to done manually. As a result, this part was dropped and is mentioned here, rather than in results.

C. Twitter

We collected tweets with the hashtags corresponding to cities we were interested in, in order to understand the sentiment of the area. For instance, if locals usually talked and or linked to news reports of criminal activity in the area, we would assume prices to be lower, as the area may be less desirable as a result.

We used the Python library Tweepy [4] to interact with the Twitter API. We tried to gather 150 tweets for each city, however, we were not able to find enough Tweets on many of the cities. Moreover, the tweets that were collected were not actionable or usable for our case. As a result, this part was dropped and is mentioned here, rather than in results.

III. DATA PROCESSING

In order for us to gain valuable insight using the data, it had to be properly processed. Firstly, we inspected the data for missing values, as is sometimes a big problem with scraping. We found that the websites were not consistent in providing data about the living area and ground area (outside area). Moreover, apartments rarely had ground area, and as such, had missing values. Missing values in ground area were handled by inserting a 0 in the ground area. Missing values in living area were not able to be handled properly, and were excluded from later analysis. There were also a substantial amount of missing values in the "Town" feature. In order to improve the quality of the data, we built a script to fill out the town from the address of the property. After applying the script to the dataset we still had 719 missing town values. These observations were labeled as being Town 0.

While inspecting the scraped data, we also noticed that the "Price" visible to us when manually navigating the websites had a discount scalar in front of them when scraped. We were therefore able to get the "Price" with the discount by using this piece of information. This is the "Price_w_Droprate" feature.

IV. METHOD

A. Annotation

For the annotation process we used the free software LabelStudio v1.6 [5]

The quantitative data we got, by scraping the data sources describes and defines the properties in an non-exhaustive manner. There is still an important aspect to consider; The state of the house. For this, we annotated the associated pictures with the property's facade to determine their state.

Due to limitations in resources, the annotation part is kept to "Proof of Concept". This means, we did a much smaller scale execution of what is possible, but provide evidence that it is possible and works. A total of 20 properties in the category 'Villa' in the area 'Copenhagen' were sampled randomly and annotated. They were evaluated on a three scale basis, with the following guideline:

Bad condition: Large areas where paint has faltered, roof has missing or replaced tiles, unkept surrounding environment
Neutral condition: Paint has faltered some, no obvious damages on the roof, tiles may have eroded some, but not missing. Surrounding environment moderately kept.
Good condition: Paint looks uniform and has not faltered, roof looks to have no obvious eroding or damage. Surrounding environment is neatly kept.

Guidelines are not perfect, but were followed strictly. However, human annotation is good for this purpose, as anything not covered under the guidelines can be reasonably dealt with. Moreover, the annotator was a Danish male aged 28 that is familiar with Danish houses and has lived in Copenhagen for 10 years. This was decided as cultural perception on Danish houses, despite guidelines, could influence the annotation decision process.

[3]https://pypi.org/project/geopy/
[4]https://www.tweepy.org/
[5]Available at: https://labelstud.io/

B. Visualizations

As we had great amounts of data that would be non intuitive to present in pure tables and numbers, we decided that focusing on visualizations to understand the data would be better. Visualizations would not only help us, readers and future users of the data set, but also provide important information regarding outliers and surprising distributions.

To visualize the data we used the Python library Plotly [6], which is an open-source plotting library. Plotly, as opposed to many other visualization libraries, has stunning visuals and many unique chart types. Using Python to create all the visualizations also greatly improves the usability for others, instead of just providing visualizations from some visualization software. As we were focused on distributions of the data we mostly constructed bar charts, donut charts and box-plots. Bar charts / donut charts were used to see how many of each type of property were present in the different areas, energy type, etc. Box-plots were used to look at the continuous data, and especially to give a clear visualization of outliers.

C. (Multiple) Linear regression

In order to understand what drivers that impact the price of properties in Copenhagen Greater Copenhagen Area, we decided upon using the linear regression model (ordinary least squares) from the Python library sci-kit learn [2]. Using a linear regression model to predict ”price” allowed us to see not only the impact of every feature, but also if it is negatively or positively impacting price. A similar approach was used before to understand the predictive power of Multiple Regression Analysis vs. Neural Network Analysis on housing data [3].

From our previous visualizations we were also able to detect outliers and wrong data insertion that could negatively affect the results of the linear regression model, and as such were excluded.

We used the built in accurary_score function to score the linear regression model, and cross-validated using the built in cross_val_score function using parametres cv=10 and scoring on r^2.

V. Results

A. Proof of Concept: Annotation Results

Although the sample size of 20 is small, more than 200 properties' pictures were looked at in the process. We found that the annotation process is only robust for properties of the type ”Villa” or ”Rækkehus” (terraced house). This is due to the fact that typically, properties of the type ”Ejerlejlighed” (owned apartments) will only have pictures of the inside. Additionally, ”Landejendom” (farm) and ”Fritidshus” (vacation home) are too different from the other types of properties that we could confidently use the guidelines to annotate their respective states.

For Villa and Rækkehus we found the annotation process to be easily followed, while providing very key information about the state of the house that could complement an over-all evaluation of the property. Moreover, we posited that being familiar with Danish homes may be important when performing the annotation.

The pictures were almost all taken in this style, where roof, facade and surrounding environment were visible. Which is required for the guidelines to apply (see Figure 1).

Fig. 1. Example of a property type Villa annotation picture.

B. Distributions in the data

First of all, we found that only 3 types may be confidently represented by the data. ”Ejerlejlighed” (apartments) were by far the most represented type with 2983 observations in the data. ”Villa” represented 2160 of observations in the data, and ”Rækkehus” (terraced house) 644 observations in the data. The last types are not excluded from the data, but we cannot make generalizations based upon them (see Figure 2).

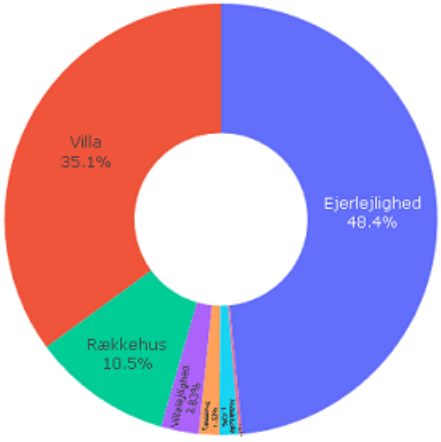

Fig. 2. Example of a property type Villa annotation picture. From: TotalHouseDistributionDonutChart on the Github

[6]https://plotly.com/python/

Based on the bias in the type of property, we wanted to investigate for bias in the Towns as well. In terms of town representation, we found that the data may be biased, as many of the properties are from Copenhagen Frederiksberg (collectively 35.52% of observations were in Copenhagen or Frederiksberg) (see Figure 3). This is important to remember when looking at further visualizations, especially price box-plots. Greater Copenhagen represents 64.48% of the data, but may have a lower price in general. As such, we expect the data to be a little biased.

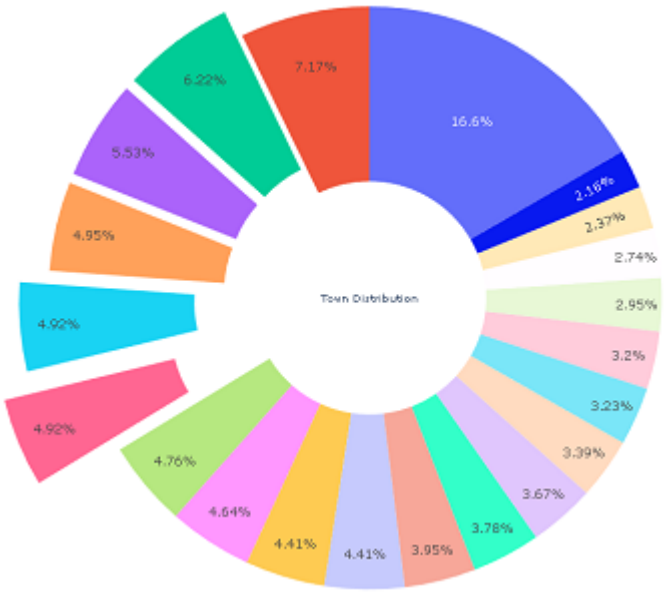

Fig. 3. Donut chart showing the distribution of towns in the data. From: TotalHouseDistributionBasedonTownDonutChart on the Github. 4 Towns with less than 1% representation were dropped from the visualization

For the price distributions we found, as expected, Villas tended to be the most expensive. Interestingly we also found that Villas are quite evenly distributed between the max and min. Moreover, the distribution has a very long tail (tail is longer, and can be seen on the Github). The apartment price distribution on the other hand does not a have long tail, instead most observations lie closer within the min, med, max in the boxplot. We hypothesized, due to later findings in the ”year built” distributions as to why the data looked this way. The idea was that Copenhagen, especially modern Copenhagen, was not built in one day. But with increasing urbanization, the city has had to expand in the last 50 years, and as a result many new apartments were built that all cost the same, depending on size (see Figure 4).

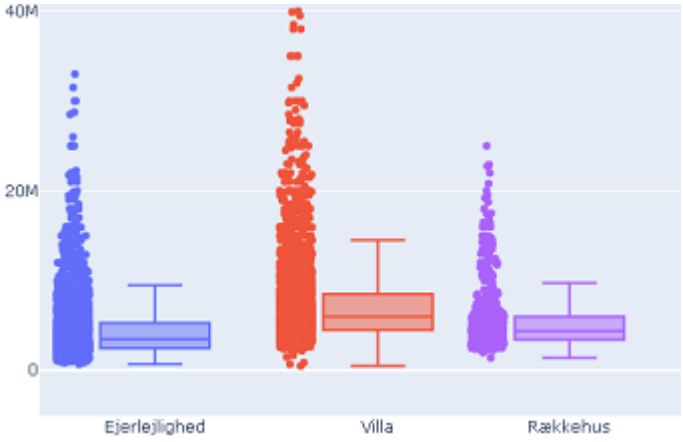

Fig. 4. Boxplot showing the spread in prices of Apartments (Blue) and Villas (Red). Type on the x-axis. Price on the y-axis. From: Price on the Github

With the last decade’s emphasis on the environment, we assumed that properties were energy friendly on average. Moreover, with the heating prices in 2022, it would’ve saved thousands upon thousands for people if their homes were energy friendly. There has been so many building projects in the last 10 years, so we expected energy label A to be very represented in the data. The energy labeling system in Denmark has 8 different categories: A, B, C, D, E, F, G. Where A is the best and G is the worst. Moreover, A has sub categories to encompass stricter energy friendliness in the last couple of years. ”A” therefore is split into ”A2010”, ”A2015”, A”2020”. We found that the combined ”A” (A10, A15, A20, A) made up only 9.47% of the properties listed. Instead, C and D were by far the most represented in the data, with respectively 36.5% and 33.6% (see Figure 5)

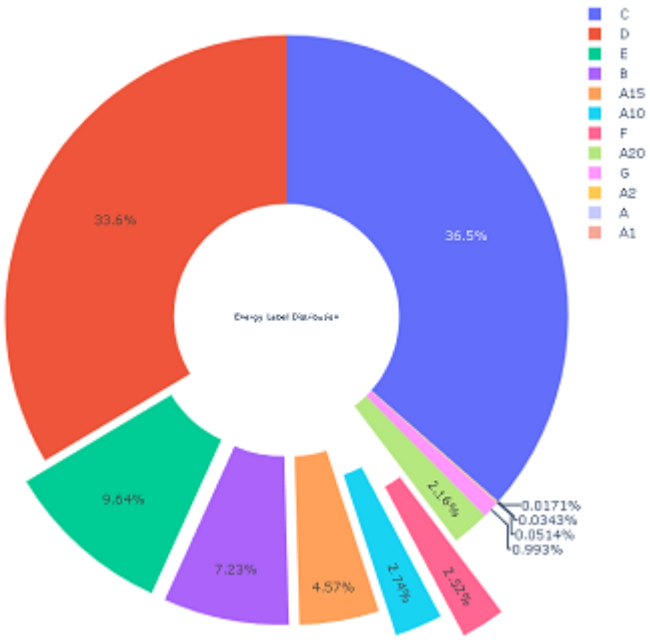

Fig. 5. Donut chart showing the different energy labels and their representation in the data. From: EnergyLabelDonutChart on the Github

As previously eluded to, we had an interesting finding when inspecting the properties and when they were built. We had assumed that a city (Copenhagen) grew gradually, and so too would housing as more people came to the city. However, there are clear clusters in the data regarding year built. We tried to visualize this using a Violin plot (see Figure 6). In doing so, we were able to show both the clusters, but also how old the different types of properties are. The clusters in the types alongside the timeline of when they were built, are interesting in the following ways: We can observe that there is no cluster (density blob) in the period around 0-20 years for villas. However, this is present both for apartments (Ejerlejlighed) and terrace houses (Rækkehus). This means that in the last century, there have been built many more apartments and rækkehuse than villas. As a result, the Villas that already existed may be more attractive due to the increase in people living there. This may in part explain why villas are so much more expensive than for instance "rækkehus".

Given the findings in the Price distribution, alongside the findings in the age of properties, we found that "Rækkehus" (terraced houses) technically are "houses/villas". However, they should be viewed as Apartments instead as they overlap completely in almost every aspect.

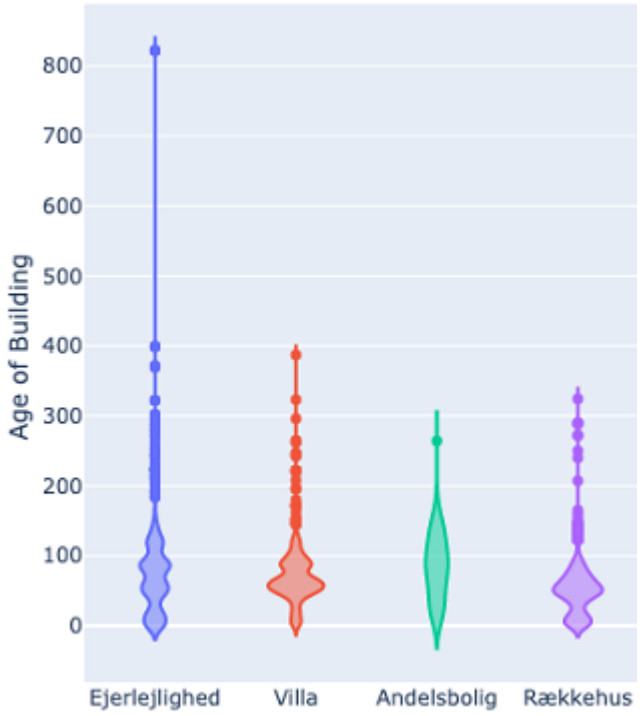

Fig. 6. Violin Plot showing the type on the x-axis and age on the y-axis. From: HouseType_Age on the Github

Lastly, as our Google Maps idea faltered, we decided to build our own interactive map of all the properties in the data using the python library "Folium" [7]. If we had the chance to gather locational data of all restaurants, institutions, etc. we could've used that instead. Moreover, we did find the map to be very useful for a person to see where properties for sale actually are located (this is not available on websites, as you would have to look up the address yourself). We found while navigating the map that the further we got to inner Copenhagen the more expensive the properties were, regardless of type (excluding non-represent types such as Farms). Figure 7 lets you see the map when it is zoomed in to Nordhavn (a part of Copenhagen). We may observe that clusters of properties indicate apartments, however, it also shows the capability of counting an object in a certain longitute/ langitute area.

Fig. 7. A snippet of Nordhavn from the interactive map, showing clusters of properties. From: Map on the Github

The map may of course be zoomed in further to reveal the individual properties, where hovering one such will reveal all the scraped information (see Figure 8 below)

[7]https://python-visualization.github.io/folium/

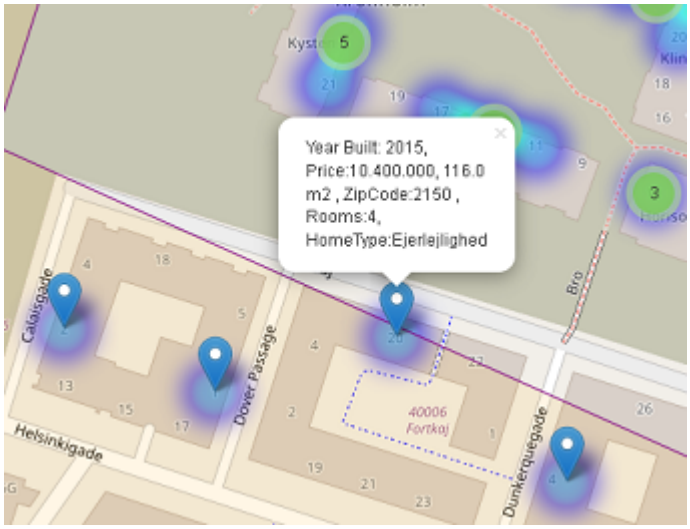

Fig. 8. A snippet of Nordhavn from the interactive map, showing the interaction of hovering a specific property. From: Map on the Github

C. Regression results

From predicting price using a linear regression model, we found the following coefficients of the results (see Appendix 1). The regression analysis in itself was not supposed to perfectly predict the price. We were instead interested in how drivers of price behaved.

We found that Energy Label (energimærke) was of great importance in predicting price. Energy label A, was associated with an increase of 400K DKK,-. However, it may be posited that if energy label A is very important, then Energy label B should be the second most positively impacting energy label. This was not found to be the case, which is discussed later.

Additionally, we discovered that several cities/towns greatly impacted the price of a property, both positively and negatively. Especially Kokkedal, Søborg, Dyssegård and Hellerup were associated with a positive impact, whereas København V, København N, Frederiksberg, Valby and Skovlunde negatively impacted price. Frederiksberg being negatively associated with price surprised us a lot, which will be discussed later.

Lastly, we could observe that all other factors, excluding energy labels and location (Towns), were not significant to the prediction in the price of houses for our data set.

VI. Discussion and Reflections

A. Data Gathering

Boliga.dk and Boligsiden.dk are both great sources of data when it comes to houses in Copenhagen. However, in hindsight, we could have used more sites, such as ”Nybolig.dk” and ”Boligportal.dk” to get additional data, as well as different data. It became apparent to us that we missed rented apartments. This was in part due to lack of time, but also we did not know how to deal with the fact that the price of the apartment was not listed, only the monthly rental expenses.

It would have been interesting if we were able to get the current price of all houses, not just the ones for sale. The problem was that it was hard to get prices for properties that had not been for sale for a long time, and previous records may not reflect the current price. For instance, we posit that the data on Frederiksberg (a Town in Denmark) may not be fully representable. It is supposed to be a very expensive part of Denmark, albeit hosting some clusters of apartments that are cheap. However, we found Frederiksberg to be negatively associated with an increase in price. This is because the properties that were scraped in large very cheaper ones. Being more aware of these occurrences may have improved the quality of our findings from the regression analysis.

B. Annotation

The annotation part of the research was as stated kept to a Proof of Concept as we failed to see a good use case in time. It is obvious that expanding on this, and incorporating it into our analysis, especially the regression, was to be preferred. Moreover, we acknowledge that the websites are trying to sell a property, and as such provide the most flattering pictures to the viewer. We don’t believe there’s a good way to perfectly correct for said bias, however, increasing the scale to have 5 choices may help better annotate. The distribtution could then be normalized, to adjust for the posited bias (perceived picture state better than actual state).

We could also have taken a different approach and instead had annotated them after how pleasant they looked. For instance, a house could be ”hyggeligt” (Danish cultural feeling best described by ”cozy”), it could be scary, etc. These could then be annotated whether or not the house’s appearance is perceived positively (hyggeligt), neutrally or negatively (scary). This however, would require a much bigger sample size and people resources to gain valuable insights.

Lastly, we acknowledge that even if the annotation was carried out in full, no annotator would have been an expert. Although we accounted for the cultural differences, the annotator lacked crucial domain knowledge to properly assess the state of the facade. A professional may know exactly what to look for, whereas we looked for the most obvious signs.

C. Visualization

For the visualizations we had so much information that it was hard to understand exactly what to report in the results section. For instance, we were not able to report and show visually that we could detect wrong data insertions on the scraped websites (We found multiple apartments that were reported as being 800 years old, but when looked up in a register to be built in 1960). Perhaps having a small section about this was preferred, although for the research question it was insignificant, and as such is reported here instead.

Additionally, some of the more advanced visualizations were not able to be reported, as they could not be fit into the double column format comfortably. Having 20 visualizations in the appendices and constantly referring the reader, would not make for a good experience. Most importantly, this hurts the reporting of the interactive map that can only be reported statically, and poorly as such. We therefore encourage the reader to use on the "Map.html" on Github [8]

D. Regression

There are a set of improvements that could be made to this section. Firstly, we did remove variables that we saw were correlated with price, however, after additional review, we see that rooms and size are still both in the regression

Secondly, the usage of one-hot encoding Towns in order to find out whether particular locations were positively or negatively associated with price may be too simplistic. It is mentioned that linear regression models run best on continuous data. Instead, we could have used a classifier model, for instance, logistic regression to better fit the data, if we could find a price-related categorical variable

Thirdly, it may have improved the quality of the results if we ran the regression separately on Types of properties. It might be that for Villas, energy label means a lot, whereas in apartments it does not, because they are all new anyways. For instance, energy label D positively impacting price is more likely than not a result of Villas (less energy friendly) are priced higher. It does not mean that having energy label D is preferred to energy label B, but could be interpreted as such. Despite these short comings, the "direction" of the results should be the same.

VII. Conclusion

We set out to answer the following question:

What are the main drivers of property prices in Copenhagen Greater Copenhagen area?

We concluded that energy label A was of key importance when determining the value of a property. This also makes sense logically as it is cheaper in the long run, and as such should be more attractive. It may also be posited that energy label A captures more information than just the energy label, and as such is of bigger importance than if it was decomposed. Nevertheless, it is the most important individual driver.

Secondly we concluded that location is an important driver of price. Although the exact impact is not perfectly captured by the model, as mentioned in the "Discussion and Reflections" section. The sentiment, however, remains and can be determined by visualizations on the Github [9].

Thirdly, we conclude that annotating the state of the property (preferably by a Danish expert) should be pursued (for villa and rækkehus only) as it provides valuable information for prediction models that was previously unobtainable.

Lastly, we concluded that using Twitter to get a sentiment of an area is not feasible. Instead we propose as a reflection that annotation can solve this problem.

The reference list is kept short as only scientific articles and libraries are referenced here. All other websites, web-articles, libraries are shown as footnotes instead.

References

[1] Sirisuriya, SCM. A comparative study on web scraping. Proceedings of 8th International Research Conference. November 2015.

[2] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. et al. "Sci-kit learn". Journal of Machine Learning Research. vol. 12, pp. 2825–2830, 2011.

[3] Nguyen, N. and Cripps, A. Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. Journal of Real Estate Research. vol. 22, pp. 313–336, 2020.

[8] https://github.com/ResitKadir1/Data_in_the_Wild/tree/master/Graphs

[9] https://github.com/ResitKadir1/Data_in_the_Wild/blob/master/sGraphs/Town_Price.png

Appendix 1. Feature Importance